評釈：PatRe ― 特許審査の全工程を対象とした Office Action・意見書生成ベンチマーク

― 論文内容の整理、批評的評価、および外部評価・評判の調査 ―

対象論文：Wang et al., "PatRe: A Full-Stage Office Action and Rebuttal Generation Benchmark for Patent Examination", arXiv:2605.03571v1（2026 年 5 月 5 日）

2026 年 9 月 12 日

Claude Fable 5.1

要旨

- PatRe は、特許審査を審査官の Office Action（OA）と出願人の意見書（rebuttal）が往復する多ターン生成タスクとして定式化した先駆的ベンチマークである。OA-DP／RO／RS という情報提供水準の段階設計、条文別の誤り分析、引用精度（RCA）の定量化、評価指標と人間評価の相関報告など、診断的価値の高い設計を備える [1]。
- しかし、収録 480 件の全てが「最終的に特許査定に至った案件」であるという生存者バイアス、明細書・図面を入力に含めない設計、自明なベースラインの欠如、LLM-as-a-judge と人間評価の乖離という 4 点が、主要な定量的結論（特に「AI 審査官は厳格すぎる」というバイアス主張とモデル間順位）の頑健性を損なっている。
- 外部評価は 2026 年 9 月時点でほぼ皆無である。被引用は実質ゼロ、学会採択は確認できず（査読前プレプリント）、データセットは未公開（GitHub はコードのみ）で、専門メディア・SNS での言及も確認できない [2,13]。実務上の教訓は論文自身の結論と一致し、現状の LLM は独立した審査・意見書システムには不十分であり、人間（弁理士・審査官）の監督が不可欠である。

1. 書誌情報と論文の位置づけ

【確認済み】 対象論文は "PatRe: A Full-Stage Office Action and Rebuttal Generation Benchmark for Patent Examination"（arXiv:2605.03571v1 [cs.CL]、2026 年 5 月 5 日投稿）であ

る[1]。著者は Qiyao Wang（中国科学院深圳先進技術研究院〔SIAT〕／中国科学院大学）、Xinyi Chen（大連理工大学）ら 7 名で、責任著者は Yuan Lin（大連理工大学）と Min Yang（SIAT）である。プロジェクトページ（patre.wangqiyao.me）と GitHub リポジトリ（AIforIP/PatRe）が公開されている[1,2]。

【確認済み】 著者グループは知財×LLM の一連の研究を展開しており、IPEval（2024 年）、AutoPatent（2024 年）、IPBench（ACL 2026 Findings 採択）を先行して発表している[3,4,5]。筆頭著者 Qiyao Wang は中国科学院大学の NLP 博士課程学生である[6]。本論文のヘッダには「IP Intelligence 2026-05」の表記があり、同グループの「IP Intelligence」シリーズの一環と位置づけられる。

【確認済み】 論文の中核的主張は、特許審査を分類・静的抽出ではなく、OA と rebuttal が往復する多ターンの動的プロセスと捉え、その全工程（full-stage）をモデル化した「初の」ベンチマークである、という点にある[1]。Table 1 では、HUPD（2023 年）、IPEval（2024 年）、IPBench（2025 年）、PILOT-Bench（2025 年）、PANORAMA（2025 年）、PEDANTIC（2025 年）を判別型、PatentEdits（2024 年）・Patent-CR（2025 年）を生成型（Full-stage は Partial）と整理し、PatRe のみが Statute／Evolution／Adversarial／Full-stage の全項目を満たすとする[1]。

2. 論文内容の整理

2.1 タスク設計

【確認済み】 2 つのタスクから構成される[1]。

- **Task 1：OA 生成（審査官役）**　3 設定を置く。(a) OA-DP（Directly Prompting：先行技術を与えず内部知識のみで判断）、(b) OA-RO（Reference-Oracle：審査官引用文献と出願人提出文献のオラクル集合を与え、モデルが選択して法的根拠を構築）、(c) OA-RS（Retrieval-Simulated：BM25 検索の上位文献と正解文献を混ぜたノイズ混じり候補プール）。GitHub の README によれば、BM25 上位 5 件と審査官引用文献を統合・重複除去して最大 10 件の候補プールを構成し、検索対象は各 OA 日付より前の文献に時間的に限定される[2]。2 ターン目以降は直前の rebuttal と審査履歴も入力される。

- **Task 2：Rebuttal 生成（出願人代理人役）**　現クレーム、OA、参照文献の要約、審査履歴を入力とし、手続的書類ではなく実質的な反論の生成を求める。出力構造には補正クレーム（REFINED CLAIMS）が含まれる [1]。

【確認済み】 プロンプト（付録 E）では、審査官役に「拒絶を前提にしない（No Bias）」「適用しない条文の見出しを書かない（Active Statute Gating）」等を課し、代理人役には「原明細書に裏付けのない補正を禁じる（No-New-Matter Rule）」を課している。ただし入力欄はクレーム、OA、参照文献要約、審査履歴のみであり、明細書本文・図面は与えられていない [1]。

2.2 評価指標

【確認済み】 客観指標として Decision Accuracy（OA 処分ラベルの一致）、Statute Precision（|S_pred∩S_gt|／|S_pred|、精度のみ）、ROUGE-L を用いる。LLM-as-a-judge には Gemini-3.1-Flash-Lite（temperature 0）を「特許監査人」として用い、Soundness／Clarity／Constructiveness／Completeness／Language Style の 5 次元を 1〜10 点で採点する。Rebuttal には各拒絶ポイントへの応答率（Point-wise Coverage）を加える [1]。OA 判定プロンプトは「処分ラベル（Non-Final／Final／Allowance）が正解と不一致なら Soundness の上限を 3 に固定する」と規定している（付録 E、Figure 10）[1]。

【確認済み】 人間評価（付録 C）は知財専攻の博士課程学生 3 名が、7 モデルから抽出した OA 100 件・rebuttal 100 件をブラインドで採点した。評価者間一致は平均スコアで Pearson 0.7285、Spearman 0.7715、Kendall 0.5861（Language Style は低く、Pearson 0.4676）。指標と人間判断の相関（Table 5）は、LLM-as-a-judge が Pearson 0.6808、Decision Accuracy が 0.4863、Statute Precision が 0.2766、ROUGE-L が 0.1672 であった [1]。

2.3 データ

【確認済み】 USPTO 公開データ（data.uspto.gov）から審査履歴を縦断的に収集した 480 件で、IPC 8 セクション（A〜H）を各 12〜13%でほぼ均等に配分している。付録 A は「全件が Notice of Allowance（NOA）後の 2024 年以降に公開された」と明記しており、収録案件は全て最終的に特許査定に至ったものである [1]。ラウンド数は OA 平均 2.24 回、rebuttal 平均 1.24 回（最大 15 回）、文書長は OA 平均 2,816 トークン、rebuttal 平均 1,836 トークンである。

【確認済み】 拒絶理由分布（Table 6、n=1,024）は §103 が 40.53%、§112 が 22.46%、

§102 が 19.73%、二重特許が 11.91%、§101 が 5.37%。OA 種別分布（Table 7、n=1,075）は Notice of Allowance が 44.56%、Non-Final Rejection が 40.47%、Final Rejection が 14.14%、Ex Parte Quayle が 0.84%である。引用文献統計（Table 8）は全体で平均引用 8.60 件・平均議論 6.30 件だが、セクション C のみ引用 2.88 件に対し議論 5.77 件（比率 200%）という不整合な値を示す[1]。

2.4 評価対象モデルと主要結果

【確認済み】 プロプライエタリ側は GPT-5-mini、GPT-4o-mini、Gemini-2.5-Flash、DeepSeek-V3.2（API 経由のため便宜上この区分）、オープン側は LLaMA3.1-8B、Qwen3.5-9B／27B、Gemma3-12B／27B、LLaMA3.3-70B を評価した。API コストは GPT-5-mini が 37.85 ドル、Gemini-2.5-Flash が 14.13 ドル等である。Claude 系、GPT-5（フル）、Gemini Pro 級のフロンティアモデルは含まれていない[1]。主要結果を表 1 に示す。

表 1　主要結果の抜粋（Decision Accuracy は%、判定スコアは 1〜10 点。出典：論文 Table 2、3、4、12）

モデル	Decision Acc.（DP／RO／RS）	OA-RO LLM 判定平均	Rebuttal LLM 判定平均	人間評価 OA-RO 平均	人間評価 Rebuttal 平均
GPT-5-mini	51.4／50.0／52.7	4.89	9.18	3.59	6.85
Gemini-2.5-Flash	50.0／46.4／52.8	4.36	8.34	3.68	6.67
DeepSeek-V3.2	47.6／49.7／42.2	4.34	8.37	4.33	5.05
GPT-4o-mini	24.4／43.3／36.3	3.61	4.50	2.97	4.80
Qwen3.5-27B	48.8／47.6／50.5	4.37	8.29	—	—
Gemma3-27B-it	39.5／45.4／38.1	3.65	5.47	3.14	5.96
LLaMA3.3-70B-it	10.2／9.7／21.8	3.40	4.59	—	—

注：人間評価は Qwen3.5 系と LLaMA3.3-70B を対象に含まない。

【確認済み】 OA 生成では 5 次元のうち Soundness（1.8〜3.4）、Constructiveness（1.2〜3.1）、Completeness（1.7〜3.9）が極端に低く、Clarity（5.3〜7.9）、Language Style（5.1〜8.7）は高い。Rebuttal 生成では GPT-5-mini が Point-wise Coverage 90.5%、Soundness 8.71、平均 9.18 で最高であった[1]。

2.5 論文の主要な知見（5 つの Finding）

- **Finding 1**　LLM は受動的防御（rebuttal）に長け、能動的な問題発見（OA）に弱い。DeepSeek-V3.2 や Qwen3.5-27B は OA では平均 4.5 未満だが rebuttal では 8 以上を示す。
- **Finding 2**　過剰批判バイアス。特許査定されるべき案件で拒絶を捏造する。真の査定案件を拒絶と予測する率は GPT-5-mini が 0.83、DeepSeek-V3.2 が 0.95、LLaMA3.3-70B が 0.93（Gemini-2.5-Flash は 0.38）。LLaMA3.3-70B は Statute Precision 54.7%にもかかわらず Decision Accuracy は 9.7%で、誤りの 37.47%が Non-Final を Final と早期判定したもの。
- **Finding 3**　条文適用の論理的不整合。§101 は捏造（FP）72.8%・見落とし（FN）48.8%の「二重失敗」、§102 は FP 47.5%・FN 40.3%、§103 は FN 7.6%・FP 47.6%、§112 は FN 19.9%・FP 60.0%の「過剰執行」。
- **Finding 4**　引用精度（RCA）は外部証拠の質に強く依存する。GPT-5-mini は DP 5.1%、RO 74.3%、RS 62.3%で、内部知識のみでは引用を幻覚する。
- **Finding 5**　ROUGE-L 等の語彙的指標は法的妥当性と乖離する（ROUGE-L と Decision Accuracy の相関は Kendall τ =0.0258）。

結論として、論文は「現状の LLM は独立した審査システムとしては不十分」と述べ、他法域・多言語への拡張を今後の課題としている [1]。

3. 評釈（批評的評価）

3.1 貢献として評価できる点

【分析】 第一に、多ターンの審査履歴を実データで再構築し、審査を「生成タスク」として定式化した点は、判別型が主流だった先行研究（HUPD、PANORAMA、PEDANTIC、PILOT-Bench）に対する明確な前進である。PANORAMA が査定経緯を保持しつつ判別型タスクに落とし込んだのに対し [7]、PatRe は OA・rebuttal 双方の「生成」を課す。第二に、OA-DP／RO／RS という情報提供水準の段階設計は、内部知識と外部証拠の寄与を切り分ける巧みな実験設計である。第三に、条文別の FP／FN 分析、RCA による幻覚引用の定量化、指標と人間評価の相関報告は、単なるスコア競争を超えた診断的価値を持つ。第四に、評価コードを公開している [2]。

3.2 方法論上の問題点

A. 生存者バイアス（最重要）【確認済み・分析】　付録 A が明記するとおり、全 480 件が NOA 後に公開された、すなわち最終的に特許査定された案件である [1]。このため、(i) 正解 OA 分布が Notice of Allowance 44.56%と査定寄りに偏る、(ii) 正解 rebuttal は事実上「最終的に成功した反論」のみとなり、rebuttal タスクの見かけの容易さと高スコア（GPT-5-mini 平均 9.18）の一因になり得る、(iii) 最終的に登録された母集団では「本来拒絶されるべき案件」が構造的に少なく、Finding 2「審査官として厳しすぎる」の解釈にも影響する。対照的に PANORAMA は登録案件と非登録案件の双方を含む 8,143 件を収録しており [7]、母集団設計の面で PatRe は弱い。

B. 自明なベースラインの欠如【分析】　OA 種別は NOA 44.56%と Non-Final 40.47%で上位 2 クラスが 85%を占める [1]。最良モデルの Decision Accuracy 52.7〜52.8%は、常に多数クラス（NOA）を予測する単純ベースライン（約 44.6%）を 1 割弱しか上回らない。論文は多数クラスやランダムのベースラインを報告しておらず、モデルの実質的な弁別力が過大に見える。

C. Soundness と処分一致の機械的連動【確認済み・分析】　判定プロンプトは処分不一致なら Soundness を 3 以下に固定する [1]。Decision Accuracy が約 50%であれば半数が機械的に 3 以下となり、Soundness 2〜3 という低スコアの相当部分は「推論の質」ではなく「処分ラベルの不一致」を反映している。Finding 2 の「形式は良いが論理が弱い」という因果解釈には留保が必要である。

D. 明細書・図面・全文先行技術の不在【確認済み・分析】　入力はクレームと参照文献の要約のみで、明細書本文・図面は与えられていない [1,2]。§112（記載要件・実施可能要件・明確性）の判断は明細書なしには原理的に困難であり、§112 の FP 60%（過剰執行）は入力設計の産物である可能性が高い。実際、ケーススタディの OA-DP 例（Table 15）では、明細書が与えられていないにもかかわらず「明細書に十分な裏付けがない」として§112(b)拒絶が生成され、さらに 2 ターン目の出力が 1 ターン目とほぼ同一（クレーム 1〜12 が削除され 13〜29 が係属しているのに「Claims 1-10 are rejected」）という不整合も見られるが、本文では論じられていない [1]。同様に、要約ベースでは厳密な claim-to-art mapping は不可能で、「モデルは表層的なマッピングしかできない」という結論の一部は入力制約に由来し得る。代理人役の No-New-Matter Rule や判定側の specification_support 基準も、明細書なしでは検証不能である。

E. LLM-as-a-judge の妥当性【確認済み・分析】 (i) 判定者 Gemini-3.1-Flash-Lite は軽量モデルであり、被評価モデルより弱い可能性がある。(ii) LLM 判定の絶対値は人間評価より大幅に高い（GPT-5-mini の rebuttal は 9.18 に対し人間 6.85、DeepSeek-V3.2 は 8.37 に対し 5.05）。(iii) 順位も食い違う。OA-RO の人間評価では DeepSeek-V3.2 が最高（4.33）で GPT-5-mini（3.59）は Gemini（3.68）を下回り、rebuttal の人間評価では Gemma3-27B（5.96）が DeepSeek-V3.2（5.05）を上回る（LLM 判定では 8.37 対 5.47 で逆）。Observation 1「GPT-5-mini が一貫して最良」は人間評価では支持されない。(iv) 人間評価は Qwen3.5 系（最良級のオープンモデル）と LLaMA3.3-70B を含まず、「プロプライエタリ対オープンの差」は人間評価で検証されていない。(v) 人間評価者は審査官・弁理士ではなく博士課程学生 3 名で、各 100 件にとどまる [1]。

F. Rebuttal 評価は「網羅性」であって「説得力・成功」ではない【分析】 Point-wise Coverage 90.5%は各拒絶ポイントに触れたか否かであり、審査官が実際に拒絶を撤回するかという結果ベースの検証はない。生存者バイアス（A）と相まって、「LLM は説得的弁護に優れる」という結論は形式指標に依存している。実務上は、AI が起草した意見書への過信リスクを示唆する。

G. 指標設計【分析】 Statute Precision は recall を報告しないため、保守的に条文を絞るモデルが有利になる。Decision Accuracy は Non-Final／Final という手続的区分を含み、実体判断（特許性）と手続判断が混在する。

H. モデル選定の非対称性と陳腐化【確認済み・分析】 被評価モデルは GPT-5-mini・Gemini-2.5-Flash 等の軽量級が中心で、GPT-5 フル、Claude 系、Gemini Pro 級は未評価である一方、判定者のみ Gemini 3.1 世代という非対称がある [1]。評価対象の Qwen3.5-27B は 2026 年 2 月公開で投稿時点では最新であったが、その後 Qwen3.6〜3.8 が相次いで公開されており [8]、2026 年 9 月時点では既に世代が進んでいる。

I. データ品質と「初」の主張【確認済み・分析】 Table 8 セクション C の「議論文献数が引用文献数を上回る（比率 200%）」という値は、抽出・解析パイプラインの不備を示唆する。「37.47%」等の過剰な有効数字、Table 15 のキャプション誤記（OA-DP の例を「OA-RO」と表記）も見られる [1]。「初の full-stage ベンチマーク」との主張については、PANORAMA（決定経緯と理由を収録するが判別型）[7] や PEDANTIC（§112(b)特化の抽出）[9] との差分を精査した限り、「OA・rebuttal 双方を多ターン生成として課す」点での新規性は妥当と評価できる。

3.3 結果解釈の妥当性

【分析】 論文の主張のうちデータが確実に支持するのは、「語彙的指標は法的妥当性と乖離する」（Finding 5）と「引用精度は外部証拠に強く依存する」（Finding 4）である。一方、Finding 2 の過剰批判バイアスは生存者バイアスと交絡しており、「LLM が本質的に厳格すぎる」のか「テストセットが査定に偏っている」のかを本データだけでは分離できない。Observation 1（GPT-5-mini 最良）は自動指標では支持されるが人間評価では覆るため、断定は適切でない。総じて、定性的結論の方向性（OA に弱く rebuttal に強い、形式は流暢だが論理は脆弱）は妥当だが、モデル間順位と厳格化バイアスの因果には過剰な主張がある。

3.4 実務への含意

- **中間処理・意見書起草**　LLM は意見書の形式的網羅性で高スコアを示すが、これは「各拒絶理由に触れた」水準であり、拒絶撤回の成否を保証しない。AI 起草の意見書は弁理士による実質的な点検が不可欠である。
- **自社出願の模擬審査（AI 審査官）用途**　過剰批判バイアスにより、本来登録され得るクレームにも拒絶を捏造する傾向がある。社内スクリーニングに用いると保守的すぎる偽陽性が多発し得る。
- **特許庁の審査 AI 導入との関係**　USPTO は任意パイロット「AI Search Automated Pilot（ASAP!）」を 2025 年 10 月 20 日に開始した（申立てと手数料を要する任意プログラムであり必須化ではない）[10]。EPO は OA の事前起草支援を検討しつつも、審査官の実質判断を支援する human-centric の方針を維持している [11]。JPO も AI アクションプラン（2022～2026 年度）で生成 AI による審査支援を検討中である [12]。本論文は、これらが「独立した審査システム」には未達で人間の監督が必須であることを実証的に補強する。
- **LLM-as-a-judge の限界**　自動判定は人間評価と順位が食い違い得る。社内で LLM 評価基盤を運用する場合も、人間評価との相関検証を省略すべきでない。

3.5 日本実務（JPO・弁理士実務）への示唆と限界

【分析】 PatRe は USPTO（35 U.S.C.、MPEP）専用であり、拒絶種別（Non-Final／Final、Ex Parte Quayle）や条文体系は JPO 実務（拒絶理由通知・意見書・補正書、進歩性の論理付け、サポート要件・実施可能要件）と直接対応しない。スコアの直接転用はできない。他方、

「審査官役では厳格化バイアス、代理人役では形式的に流暢」という傾向は、日本語 LLM を意見書ドラフトや模擬審査に用いる場合にも一般化する可能性が高い。特にサポート要件（特許法 36 条 6 項 1 号）・明確性要件（同 2 号）の判断には明細書全文が不可欠であり、「明細書不在では§112 判断が破綻する」という本論文から読み取れる知見は、JPO 文脈での入力設計（明細書・図面を必ず与える）への直接的な教訓となる。

4. 外部評価・評判の調査結果

【確認済み】 2026 年 9 月 12 日時点で、PatRe に対する外部評価は極めて限定的である。

- **学術的反応・引用**　Hugging Face の論文ページは本論文を引用するモデル・データセット・Space がいずれも 0 件と表示しており [13]、被引用は実質ゼロである。後続研究で PatRe を引用・利用したものは確認できなかった。Semantic Scholar・OpenAlex の正確な被引用数は API 制限のため直接確認できていない（未検証）。
- **学会採択**　arXiv のコメント欄に採択情報はなく、ACL Anthology 等にも登録がない。査読前プレプリントの段階にとどまる [1]。
- **コード・データ**　GitHub は 9 スター・0 フォーク・70 コミットでリリースはなく、公開されているのはコードのみである [2]。README のタイムラインは「Code」のみ完了で「Dataset」「Software Demonstration」は未完、データセットのリンクはプレースホルダのままである。抄録は「コードとデータセットを公開する」と述べるが、実際にはデータセットは未公開であり、Hugging Face 上にも該当データセットは存在しない [13]。したがって第三者による追試は現時点で不可能である。
- **SNS・ブログ・専門メディア**　YouTube に自動生成的な解説動画が 1 本あるのみで [14]、Reddit、Hacker News、X、LinkedIn、および特許専門メディア（IPWatchdog、Patently-O、IAM、Managing IP）に実質的な言及は見当たらない。中国語圏（机器之心、量子位、PaperWeekly、知乎）・日本語圏でも取り上げは確認できなかった。Hugging Face 上の議論は著者自身の告知と自動投稿のみである [13]。

【推論】 反応の乏しさの主因は、(i) 投稿から約 4 か月と日が浅いこと、(ii) データセット未公開で追試・利用のハードルが高いこと、(iii) 同時期に競合ベンチマーク（PANORAMA〔NeurIPS 2025 採択〕[7]、PILOT-Bench〔NLLP 2025〕[15]、PEDANTIC

〔PatentSemTech@SIGIR 2025〕[9]）が存在し注目が分散していること、と考えられる。

5. 論文が引用する USPTO 統計の検証

【確認済み】 論文導入部が引用する USPTO の 2025 年統計を一次情報に照合した。

- **バックログ 837,928 件**　正確。USPTO 長官 Squires の 2025 年 AIPLA 年次総会講演は、未審査案件が 2020 年の 576,103 件から 2025 年 1 月の 837,928 件に増加したと述べている[16]。なおその後は減少に転じ、2026 年 4 月 6 日時点では 776,995 件である[17]。
- **一次審査待ち期間（first-action pendency）20.5 か月**　USPTO のブログは FY2023 の値として 20.5 か月、FY2024 は 19.9 か月と記載する[18]。USPTO の議会向け提出資料は、2025 年 1 月の 20.5 か月から 22.2 か月へ上昇したと明記しており[19]、論文が「2025 年に 20.5 か月」と記す点は時点の取り方にずれがある【分析】。
- **出願件数 475,223 件**　USPTO の議会向け提出資料は FY2025 の新規特許出願（連番付与ベース）を約 475,233 件と記載しており[19]、論文の「475,223 件」は下 3 桁が異なる。論文側の転記誤記の可能性が高い【分析】。

6. 総合評価

【分析】 PatRe は、特許審査を多ターン生成タスクとして定式化した先駆的で野心的なベンチマークであり、OA-DP／RO／RS の段階設計、条文別の FP／FN 分析、RCA、指標と人間評価の相関報告など、診断的価値の高い設計を持つ。しかし、(A) 全件が特許査定案件という生存者バイアス、(D) 明細書・図面の不在、(B) 自明なベースラインの欠如、(E) LLM 判定と人間評価の乖離という 4 点は、主要な定量的結論（特に厳格化バイアスとモデル順位）の頑健性を損なう。データセット未公開のため追試も困難で、外部の学術的・実務的反応は現時点でほぼ皆無である。

実務家にとっての最大の教訓は、論文自身の結論どおり「現状の LLM は独立した審査・意見書システムには不十分で、人間の監督が不可欠」という点にある。日本実務にスコアは直接転用できないが、「AI 審査官は保守的に拒絶を捏造しやすい」「AI 意見書は形式が流暢でも実質の検証が必須」「§112・サポート要件の判断には明細書全文が必須」という傾向は、普遍的な示唆として受け止めるべきである。

参考文献

[1] Wang, Q., Chen, X., Chen, L., Wang, H., Alinejad-Rokny, H., Lin, Y., Yang, M., "PatRe: A Full-Stage Office Action and Rebuttal Generation Benchmark for Patent Examination," arXiv:2605.03571v1 [cs.CL], 2026 年 5 月 5 日. https://arxiv.org/abs/2605.03571（HTML 版 https://arxiv.org/html/2605.03571v1）

[2] AIforIP, "PatRe" GitHub リポジトリ（2026 年 9 月 12 日閲覧：9 スター・0 フォーク・70 コミット、コードのみ公開）. https://github.com/AIforIP/PatRe

[3] Wang, Q. et al., "IPEval: A Bilingual Intellectual Property Agency Consultation Evaluation Benchmark for Large Language Models," arXiv:2406.12386, 2024. https://arxiv.org/abs/2406.12386

[4] Wang, Q. et al., "AutoPatent: A Multi-Agent Framework for Automatic Patent Generation," arXiv:2412.09796, 2024. https://arxiv.org/abs/2412.09796

[5] Wang, Q. et al., "Towards IP Intelligence: Benchmarking Large Language Models on Intellectual Property Knowledge and Practice (IPBench)," Findings of ACL 2026. https://aclanthology.org/2026.findings-acl.1000/

[6] Qiyao Wang, 研究者プロフィール（alphaXiv）. https://www.alphaxiv.org/@qiyao-wang

[7] Lim, H. et al., "PANORAMA: A Dataset and Benchmarks Capturing Decision Trails and Rationales in Patent Examination," NeurIPS 2025 Datasets and Benchmarks Track. https://arxiv.org/abs/2510.24774

[8] Qwen Team, Qwen3.5-27B（Hugging Face、2026 年 2 月公開）. https://huggingface.co/Qwen/Qwen3.5-27B ／ QwenLM/Qwen3.8（GitHub）. https://github.com/QwenLM/Qwen3.8

[9] Knappich, V. et al., "PEDANTIC: A Dataset for the Automatic Examination of Definiteness in Patent Claims," PatentSemTech@SIGIR 2025, arXiv:2505.21342. https://arxiv.org/abs/2505.21342

[10] USPTO, "AI Search Automated Pilot (ASAP!) Program," Federal Register 90 FR 48161（2025 年 10 月 8 日告示、2025 年 10 月 20 日開始）. URL 未検証（federalregister.gov の告示ページ）

[11] "Searching for prior art in the age of AI: The EPO's approach," epi Information, Issue 3/2025. https://information.patentepi.org/issue-3-2025/the-epo-approach.html

[12] 特許庁, "Action Plan for Utilization of Artificial Intelligence (AI) Technology (Rev. 2025)". https://www.jpo.go.jp/e/system/laws/sesaku/ai_action_plan/ai_action_plan-fy2025.html

[13] Hugging Face, Paper page "PatRe: A Full-Stage Office Action and Rebuttal Generation Benchmark for Patent Examination"（2026 年 9 月 12 日閲覧）. https://huggingface.co/papers/2605.03571

[14] AI Research Roundup, "PatRe: New LLM Benchmark for Patent Prosecution"（YouTube）. https://www.youtube.com/watch?v=GSVwGDfgJ3U

[15] Jang, Y. et al., "PILOT-Bench: A Benchmark for Legal Reasoning in the Patent Domain with IRAC-

Aligned Classification Tasks," Proceedings of the Natural Legal Language Processing Workshop 2025. https://aclanthology.org/2025.nllp-1.17/

[16] USPTO, "Remarks by Director Squires at the 2025 AIPLA Annual Meeting," 2025 年 10 月 31 日. https://www.uspto.gov/about-us/news-updates/remarks-director-squires-2025-aipla-annual-meeting

[17] USPTO, "USPTO turns the corner on unexamined patent application backlog reduction," 2026 年 4 月（2026 年 4 月 6 日時点の未審査案件 776,995 件）. https://www.uspto.gov/about-us/news-updates/uspto-turns-corner-unexamined-patent-application-backlog-reduction

[18] USPTO, "Patent pendency goals: A road map for the future," USPTO Director's Blog（FY2023 first-action pendency 20.5 か月、FY2024 19.9 か月）. https://www.uspto.gov/blog/patent-pendency-goals-road-map-future

[19] USPTO, 下院司法委員会知財小委員会向け提出資料（FY2025 新規出願約 475,233 件、first-action pendency 20.5 か月→22.2 か月）. ptacts.uspto.gov 上のダウンロードリンク経由で取得（恒久 URL 未検証）